RESEARCH PAPER

Visual Entity Linking via Multi-modal Learning

Qiushuo Zheng[1], Hao Wen[2], Meng Wang[2,3] & Guilin Qi[2,3†]

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

[2]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China

[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China

Keywords: Knowledge graph; Multi-modal learning; Entity linking; Learning to rank; Knowledge graph representation

Citation: Zheng, Q.S., et al.: Visual entity linking via multi-modal learning. Data Intelligence 4(1), 1-19 (2022). doi: 10.1162/dint_a_00114

Received: September 15, 2021; Revised: November 1, 2021; Accepted: November 18, 2021

ABSTRACT

Existing visual scene understanding methods mainly focus on identifying coarse-grained concepts about the visual objects and their relationships, largely neglecting fine-grained scene understanding. In fact, many data-driven applications on the Web (e.g., news-reading and e-shopping) require accurate recognition of much less coarse concepts as entities and proper linking them to a knowledge graph (KG), which can take their performance to the next level. In light of this, in this paper, we identify a new research task: visual entity linking for fine-grained scene understanding. To accomplish the task, we first extract features of candidate entities from different modalities, i.e., visual features, textual features, and KG features. Then, we design a deep modal-attention neural network-based learning-to-rank method which aggregates all features and maps visual objects to the entities in KG. Extensive experimental results on the newly constructed dataset show that our proposed method is effective as it significantly improves the accuracy performance from 66.46% to 83.16% compared with baselines.

1. INTRODUCTION

Visual scene understanding is inevitably regarded as one of the core functions of the next-generation machine intelligence, and it has been evolving to meet increasing demands not limited to objects detection from images and video clips, more often than not, for grasping the story behind the pixels. Due to the

† Corresponding author: Guilin Qi (Email: gqi@seu.edu.cn; ORCID: 0000-0002-1957-6961).

representation power of graph structure, the scene graph [1] has been proposed to harness external knowledge for a better scene understanding. Thus parsing an image into a scene graph has attracted much research attention over the past few years. Although great efforts have been made on generating scene graphs from images, the existing methods only detect visual objects at a coarse-grained concept level (i.e., categories), which sometimes offer little help for a deeper scene understanding. In many practical scenarios, such as news-reading and e-shopping, we require entity level visual objects detection for the next question answering or recommendation systems.

Motivating example: Consider the following two scenarios: 1). An online user is reading sports news about basketball, and wants to distinguish *Yao Ming* and *Tracy Mcgrady* in the group photo, as shown in Figure 1. However, even the world-leading object recognition system cannot guarantee to provide the right answer[①]. 2). Another user tends to be interested in *Tracy Mcgrady*'s shoes and would like to know the specific signature sneaker, but the existing image search engine like *Bing.com* can only recognize white shoes. To accomplish the above tasks raised by the users, we need more auxiliary information in detail to complement the visual learning. The complementary information can be obtained from comprehensive multi-modal knowledge graphs (KG), such as DBpedia[②] and IMGpedia[③]. If the entities in KG are successfully linked to the objects in the image, we can answer the question with the right name (i.e., *Tracy Mcgrady*) in Case 1 and precisely recommend to the user in Case 2 with the specific shoe's brand (i.e., *Adidas T-MAC 4*).

Challenges: To achieve fine-grained visual entity linking in scene understanding, there are some challenges as follows: 1) It is difficult to recognize all the visual objects from an image solely based on visual information. Even state-of-the-art fine-grained object detection models trained with enormous labeled samples cannot guarantee to accurately classify all the fine-grained classes. 2) When linking entities extracted from KGs to visual objects detected from an image, the visual entity linking algorithm needs to effectively exploit heterogeneous features and utilize the cross-modal correlations to achieve disambiguation and find the accurate entity in KG for each visual object in an image.

Solutions: In this paper, we proposed a novel framework to achieve visual entity linking in visual scene understanding. Specifically, we first generated a coarse-grained scene graph for an image and extracted the visual features of the objects by employing a VGG-16 network. Then, we unitized the Gated Recurrent Unit (GRU) language method to extract textual features of objects from image caption and discover candidate KG entities by named mentions matching. After extracting the KG features for the candidate entities, we proposed a deep modal-attention neural network-based learning-to-rank method to aggregate all features and map visual objects to the entities in KG.

① Input the image to Tencent's Celebrity Detection API, Yao Ming is identified correctly, but Tracy Mcgrady is identified as Will Smith.

② https://wiki.dbpedia.org/

③ http://imgpedia.dcc.uchile.cl/

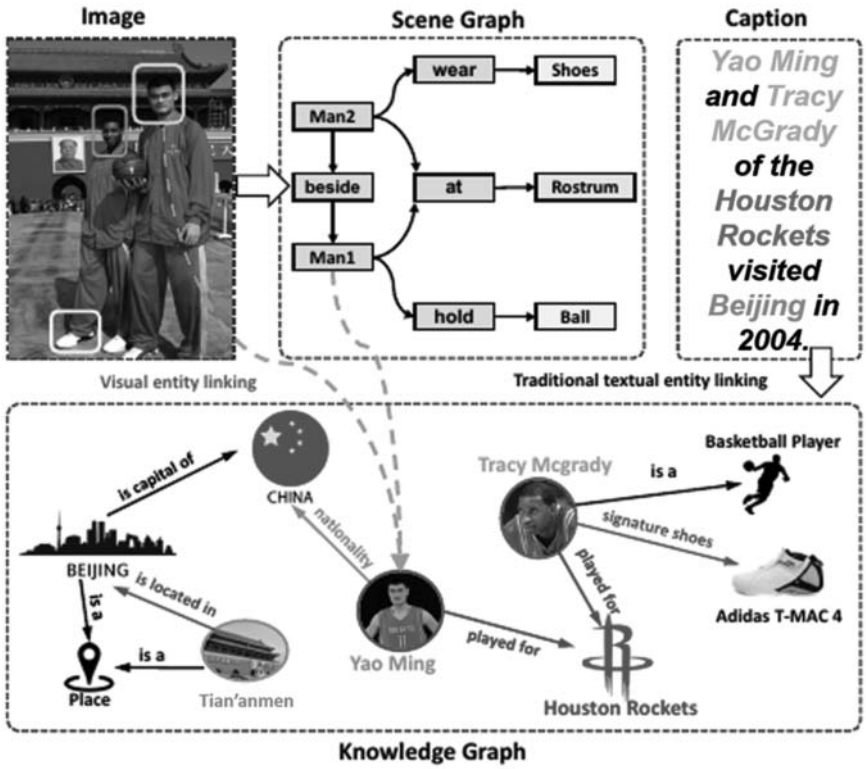

Figure 1. Scene understanding and visual entity linking[④].

Contributions: The contributions of this paper are summarized as follows:

- We are the first to consider the visual entity linking in scene graphs and have constructed a new large-scale dataset for the challenging task.
- We proposed a novel framework to learn features from three different modalities and simultaneously designed a learning-to-rank based visual entity linking model, which aggregated different features by a deep modal-attention neural network.
- We conducted extensive experiments to evaluate our visual entity linking model against state-of-the-art methods. Results on the constructed dataset show that our proposed method is effective as it significantly improved the accuracy performance from 66.46% to 83.16% compared with baselines.

2. RELATED WORK

This section discusses the existing related research from the following aspects: entity linking, visual scene understanding, and multi-modal learning.

④ Pictures were downloaded from http://sports.sina.com.cn/basketball/nba/2018-10-09/doc-ifxeuwws1965640.shtml.

2.1 Entity Linking

Entity linking is the task of mapping the mentions in the text to the corresponding entities in KG. Conventional models are usually distinguished by the supervision they require, i.e., supervised or unsupervised methods. The supervised methods [2, 3, 4, 5] utilized the annotated data to train binary classifiers or ranking models to realize entity disambiguation. The unsupervised methods [6, 7, 8, 9, 10] generally used some similarity measures between the mentions in the text and the entities in KG.

In contrast to the traditional research in the textual domain, the visual entity linking task has the following differences: (1) The form of visual representation is much more complicated than the form that may appear in textual content, and (2) Different modalities have different characteristics.

2.2 Visual Scene Understanding

Visual scene understanding includes many kinds of work, such as traditional computer vision tasks like image recognition [11, 12] and higher-level scene reasoning tasks like scene graph generation. In recent years, considerable progress has been made on many sub-issues of the overall visual scene understanding problem. Since the early work [1, 13, 14] generated the visually-grounded graph over the objects with their relationships in an image, many models have been proposed to improve the performance, such as adding prior probability distribution [13, 15, 16] and introducing the message passing mechanism [17, 18, 19].

2.3 Multi-modal Learning

Multi-modal learning [20, 21, 22] focuses on learning with contextual information from multiple modalities in a joint model. The recent relevant tasks to our work include the cross-modal entity disambiguation and entity-aware captioning. Ref. [23] built a deep zero-shot multi-modal network for social media posts disambiguation, but they are still limited to link the textual entities in the posts to the knowledge base. Refs. [24, 25, 26, 27, 28] proposed multi-modal entity-aware models to achieve image caption generation, which is also different from our visual entity linking task.

Recently, Li et al. [29] presented a multimedia knowledge extraction system, enabled the search of graph queries, and retrieved multimedia evidence. However, when dealing with visual entity linking, it adopts redundant rules for different entity types, increasing the complexity of the model.

Entity linking is an important branch in the field of natural language processing, but the existing model cannot solve the problem of multi-modal learning in entity linking. This paper proposed a joint model to provide ideas for solving this problem.

3. PROPOSED MODELS

3.1 Problem Formulation

In the following section, we will describe our multi-modal learning model for visual entity linking in detail. As illustrated in Figure 2, the proposed model includes the feature extraction module and the visual entity linking module.

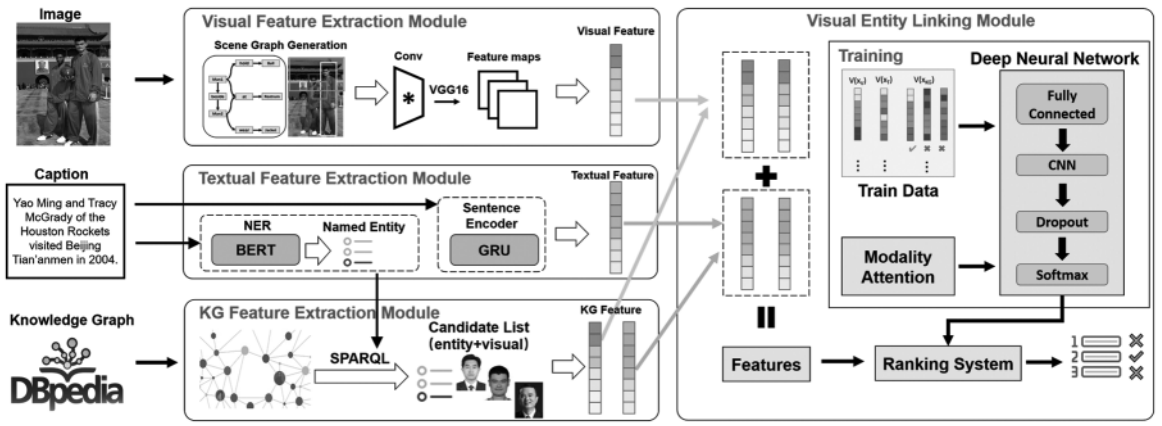

Figure 2. Overview of visual entity linking, which consists of two parts independently, namely the feature extraction module and the visual entity linking module. The feature extraction module extracts features from three modalities[5].

For the training process, given a dataset of input samples for the visual entity linking task, the number of input samples is $\mathbf{m}$, denoted by $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{m}$, with the ground truth entities $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{m}$, $\mathbf{x}_i$ is the combination of $\mathbf{v}_i^j$ and $\mathbf{t}_i^j$, and $\mathbf{y}_i$ is the total set of $\mathbf{y}_i^j$, $\mathbf{v}_i^j \in \mathbb{R}^4$ presents the j-th bounding box of the i-th image, each input sample $\mathbf{x}_i$ has a corresponding textual information $\mathbf{t}_i$ and the length L_t, and $\mathbf{y}_i^j$ represent the highest possibility entity in the multi-modal knowledge graph **KG** linked to $\mathbf{v}_i^j$. We aim to learn a transformed function $\mathbf{f}(\cdot)$ which satisfies:

$$\mathbf{y} = \underset{\mathbf{y}' \in \mathbf{KG}}{\operatorname{argmax}} \operatorname{sim}\left(\mathbf{f}(\mathbf{x}_v, \mathbf{x}_t), \mathbf{y}'\right) \tag{1}$$

where $\mathbf{f}(\cdot)$ is a transformed function that projects input samples $\mathbf{x}_v$ and $\mathbf{x}_t$ into the same space as $\mathbf{y}$, and $\operatorname{sim}(\cdot)$ generates a similarity score between prediction and the ground truth.

For the testing process, given a test sample $\mathbf{s}_p$, including the image with the corresponding caption, i.e., $\mathbf{s}_t = (\mathbf{v}_p, \mathbf{t}_p)$. The bounding boxes of test image are generated by the scene graph, represented by $\mathbf{v}_p^q \in \mathbb{R}^4$ for the q-th bounding box of the p-th image. Then, the visual entity linking task is to match the bounding box $\mathbf{v}_p^q$ to the entity $\mathbf{y}_p^q$ in **KG** by the function $\mathbf{f}(\mathbf{v}_p^q, \mathbf{t}_p)$.

[5] Pictures were downloaded from http://sports.sina.com.cn/basketball/nba/2018-10-09/doc-ifxeuwws1965640.shtml.

3.2 Feature Extraction Module

The feature extraction module aims to extract features from three modalities as follows:

Visual features: To link the image bounding boxes $\mathbf{x}_v$ to the given KG, we first used the outperforming method from [30] for the generation of the bounding boxes in the scene graph. Then, we used the VGG-16 network [31] for visual feature extraction of the image bounding boxes. The final layer representation $\mathbf{e}(\mathbf{x}_v)$ of the VGG-16 network is transformed into low dimensions, which describes the features of an image bounding box.

Textual features: To extract textual features from the image caption $\mathbf{x}_t$, we encoded the caption by a GRU language model [32] with distributed word semantics embeddings $\mathbf{e}(\mathbf{x}_t)$. We used the following implementation for the GRU.

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1}) \tag{2}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_{xr}\mathbf{x}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1}) \tag{3}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{r}_t \odot (\mathbf{W}_{hh}\mathbf{h}_{t-1})) \tag{4}$$

$$\mathbf{e}(\mathbf{x}_t) = (\mathbf{1} - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t + \mathbf{z}_t \odot \mathbf{h}_{t-1} \tag{5}$$

where $\mathbf{h}_t$ is a hidden layer output at decoding Step t, $\mathbf{r}_t$ controls the influence of the hidden layer unit $\mathbf{h}_{t-1}$ at the previous moment on the current word $\mathbf{x}_t$, $\mathbf{z}_t$ decides whether to ignore the current word $\mathbf{x}_t$, and $\mathbf{e}(\mathbf{x}_t)$ is the output textual vector representation of GRU at the last decoding step $t = T$.

Because pre-trained models can effectively represent the semantic distribution of words in a sentence, we used the pre-trained embeddings from the GloVE model [33] in the GRU sentence encoder.

Similar to the traditional entity linking models, we also need to obtain the list of named entities in the image caption $\mathbf{x}_t$. We implemented a Named Entity Recognizer (NER) based on BERT [34], Bi-LSTM, and Conditional Random Fields (CRF), as shown in Figure 3. The Bi-LSTM extracts higher-level structural information, and the CRF is used as a sequence classifier. We fine-tuned the model on our dataset and tried to achieve the best recognition results. Once we established the named entity list with the NER, we sent the list to a SPARQL query engine and obtained the candidate entities in the KG.

The structure details of our named entity recognizer are based on BERT, and it is mainly composed of BERT, Bi-LSTM and CRF.

KG features: To generate the linked candidate entities, we proposed a matching algorithm based on rules. Then, we obtained each candidate entity's image embedding $\mathbf{e}(\mathbf{y}_v)$ and structural text information embedding $\mathbf{e}(\mathbf{y}_t)$ as its KG features.

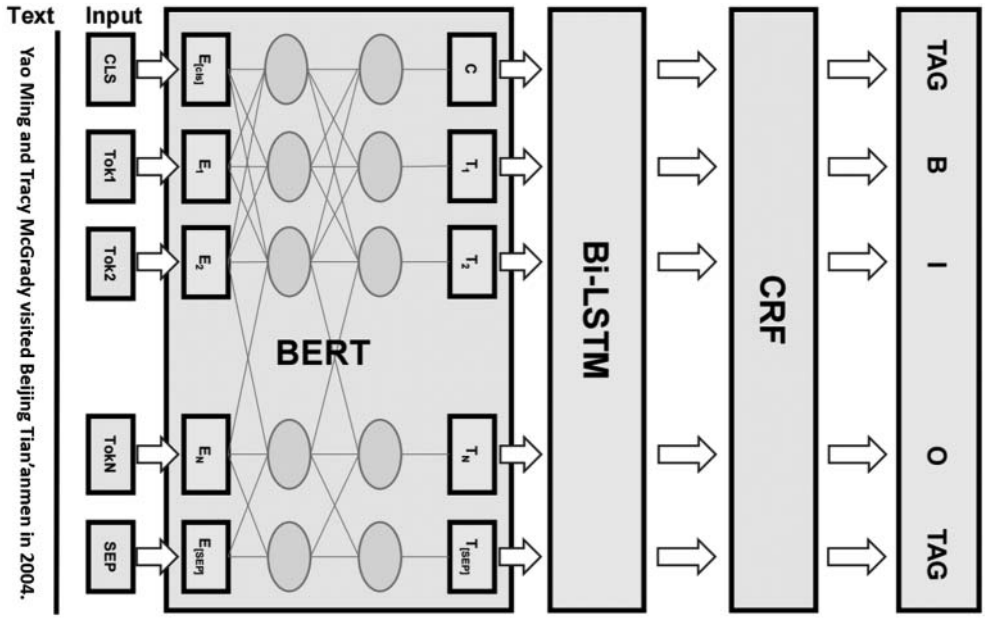

Figure 3. The structure details of our named entity recognizer based on BERT. It is mainly composed of BERT, Bi-LSTM and CRF.

Because of the inaccuracy and incompleteness of the entity mention in the caption $\mathbf{x}_t$, i.e., the abbreviations and nicknames for the person name, directly sending the entity mentions occurred in captions to the KG SPARQL query engine may not establish a complete list of candidate entities. To this end, we implemented a rule-based candidate entity list generator by using partial matching strategy and four rules as follows:

- The entity name has several words with the entity mention in common;
- The entity name is wholly contained in or contains the entity mention;
- The entity name exactly pairs the first letters of all words in the entity mention; and
- The entity name has a high string similarity over 80% with the entity mention in Levenshtein distance.

For the KG visual embedding $\mathbf{e}(\mathbf{y}_v)$ of $\mathbf{y}$, we also used the same VGG-16 network to extract the dense visual features. For the KG structural text embedding $\mathbf{e}(\mathbf{y}_t)$ of $\mathbf{y}$, we used the DBpedia [35] as our multi-modal knowledge graph to obtain the embedding with the complex model proposed by [36], which is the state-of-the-art model. The KG structural text embeddings are learned by the following score function to measure a fact <*h*, *r*, *t*> in KG:

$$f_r(h,t) = \mathrm{Re}\left(\mathbf{h}^{\top}\mathrm{diag}(\mathbf{r})\bar{\mathbf{t}}\right) = \mathrm{Re}\left(\sum_{i=0}^{d-1}[\mathbf{r}]_i \cdot [\mathbf{h}]_i \cdot [\bar{\mathbf{t}}]_i\right) \tag{6}$$

where $\overline{t}$ is the conjugate of **t** and **Re**(·) means taking the real part of a complex value. *h* and *t* are entities in KG and may be possible matched entities for **y**.

3.3 Visual Entity Linking Module

In this module, we aggregated all the three modality features to predict the best matched KG entity $\mathbf{y}_i^j$ for each image bounding box $\mathbf{v}_i^j$.

We proposed a supervised visual entity linking module using visual confidence and textual confidence. The confidence of visual similarity for the *i*-th bounding box and *j*-th candidate entity is calculated between the image feature vectors $\mathbf{e}\left(\mathbf{x}_v^i\right)$ and the visual feature $\mathbf{e}\left(\mathbf{y}_v^j\right)$ extracted from KG. The confidence of textual similarity for the *i*-th bounding box and *j*-th candidate entity is calculated between the sentence vector $\mathbf{e}\left(\mathbf{x}_t^i\right)$ and the structural text information embeddings $\mathbf{e}\left(\mathbf{y}_t^j\right)$.

$$\mathcal{L} = \sum_{s=1}^{m}\left(\lambda_t \mathcal{L}_{\mathrm{T}}\left(\overline{\mathbf{x}}^s\right) + \lambda_v \mathcal{L}_{\mathrm{V}}\left(\overline{\mathbf{x}}^s\right)\right) + \|\mathbb{W}\|_2^2 \tag{7}$$

$$\mathcal{L}_{\mathrm{T}}(\overline{\mathbf{x}}) = \sum\left[\gamma + \mathrm{conf}_t\left(\mathbf{y}'\right) - \mathrm{conf}_t(\mathbf{y})\right]_+ \tag{8}$$

$$\mathcal{L}_{\mathrm{V}}(\overline{\mathbf{x}}) = \sum\left[\gamma + \mathrm{conf}_v\left(\mathbf{y}'\right) - \mathrm{conf}_v(\mathbf{y})\right]_+ \tag{9}$$

$$\mathrm{conf}_t\left(\mathbf{y}^i\right) = \frac{\exp\left(f\left(\mathbf{e}\left(\mathbf{x}_t\right), \mathbf{e}\left(\mathbf{y}_t^i\right)\right)\right)}{\sum_{j\in\mathbb{C}}\exp\left(f\left(\mathbf{e}\left(\mathbf{x}_t\right), \mathbf{e}\left(\mathbf{y}_t^i\right)\right)\right)} \tag{10}$$

$$\mathrm{conf}_v\left(\mathbf{y}^i\right) = \mathrm{cosine}\left(\mathbf{e}\left(\mathbf{x}_v\right), \mathbf{e}\left(\mathbf{y}_v^i\right)\right) \tag{11}$$

The loss function consists of two parts, where $\mathcal{L}_{\mathrm{T}}(\cdot)$ is the supervised max-margin ranking loss for KG entity prediction on the textual features, and the $\mathcal{L}_{\mathrm{V}}(\cdot)$ is the max-margin ranking loss on the visual features. λ_t and λ_v denote hyper-parameters to tune the function. conf_v is the confidence score in the visual modality, and conf_t is the confidence score in the textual modality. Through our max-margin ranking loss function, the confidence of the correct linking entity *conf*(**y**) should be higher than that of any other candidate entity *conf*(**y′**) with the margin γ, $[\mathbf{x}]_+$ denoting the positive part of **x**.

m is the number of the samples, and *s* is the serial number of the input sample. $f(\cdot)$ is a function that projects textual embeddings into KG structural embedding space. conf_t ($\mathbf{y}^i$) is the confidence score between the caption textual embedding $\mathbf{e}(\mathbf{x}_t)$ and KG structural text embedding of *i*-th candidate entity $\mathbf{e}\left(\mathbf{y}_t^i\right)$ in textual modality, and conf_v ($\mathbf{y}^i$) is the confidence score between the image visual embedding $\mathbf{e}(\mathbf{x}_v)$ and KG visual embedding of *i*-th candidate entity $\mathbf{e}\left(\mathbf{y}_v^i\right)$ in visual modality.

To learn the different weights of modalities, we formulated the modal-attention module as follows, which selectively weakens or magnifies the different modalities:

$$[\mathbf{a}_t;\mathbf{a}_v]=\sigma\left(\mathbf{W}\cdot[\mathbf{x}_t;\mathbf{x}_v]+\mathbf{b}\right) \tag{12}$$

$$\alpha_m=\frac{\exp(\mathbf{a}_m)}{\sum_{m'\in\{t,v\}}\exp(\mathbf{a}_{m'})}\quad \forall m\in\{t,v\} \tag{13}$$

$$\bar{\mathbf{x}}=\sum_{m\in\{t,v\}}\alpha_m\mathbf{x}_m \tag{14}$$

where $\alpha=[\alpha_t;\alpha_v]\in\mathbb{R}^2$ is an attention vector, and $\bar{\mathbf{x}}$ is the final context vector that reasonably focuses on different modalities.

At test time, the following entity-predicting nearest neighbor (1-NN) classifier is used for the prediction, where λ_1 and λ_2 are the hyper-parameters:

$$1-\mathrm{NN}(\bar{\mathbf{x}})=\mathrm{argmax}\left\{\lambda_1\cdot\mathrm{conf}_t\left(\mathbf{y}^t\right)+\lambda_2\cdot\mathrm{conf}_v\left(\mathbf{y}^t\right)\right\} \tag{15}$$

$$\lambda_1+\lambda_2=1 \tag{16}$$

4. EXPERIMENTS

We constructed a new dataset for the task of visual entity linking and compared our model with state-of-the-art methods on the dataset.

4.1 Experiments Setting

Datasets. For the visual entity linking task, we need to link the image bounding boxes to specific entities in KG, which goes beyond identifying the category of the objects. However, most of the existing computer vision datasets contain no named entities in the images or captions. Therefore, we built a new dataset, namely Visual Entity Linking Dataset (VELD), which is composed of 39,000 news image and textual caption pairs with the links to KG entities, and manually labeled by expert human annotators (entity types: PER, LOC, ORG). In total, we have gathered 39,000 images with textual captions, randomly split into 31,000 for training, 4,000 for validation, and 4,000 for testing.

It is meaningless to the visual entity linking task if no word in captions describes named entities. Therefore, we performed a filtering program about named entities in VELD to remove news data that does not contain named entities, to ensure that our named entities appear in the sentence at 100%. Key aspects are summarized in Table 1. The VELD dataset, similarly to BreakingNews [37], exhibits longer average caption lengths than image-caption datasets like MSCOCO [38], indicating that news captions tend to be more descriptive.

Table 1. Statistics of datasets.

	MSCOCO	BreakingNews	VELD
Average caption length	11.30	28.09	19.68
Words related to named entities	0	15.66%	29.24%
Sentences containing named entities	0	90.79%	100%
Nouns	33.45%	55.59%	48.19%
Adjectives	27.23%	7.21%	4.16%
Verbs	10.72%	12.57%	11.49%
Pronouns	1.23%	1.36%	1.49%

Tasks. Given an image bounding box and an accompanying caption, our goal is to link the image bounding box to the corresponding KG entities in DBpedia 2016 by unitizing the visual features, textual features, and KG features.

Evaluation metrics. The primary metric of our evaluation is the accuracy of the visual entity linking to the KG entity. The accuracy is defined as in Equation (17):

$$\text{accuracy} = \frac{\left|\{\text{correctly linked entity mentions}\}\right|}{\left|\{\text{all links generated by our method}\}\right|} \tag{17}$$

Implementation details: We initialized the NER stream of our model with a BERT language model pre-trained on the English Wikipedia. Specifically, we used the $BERT_{BASE}$ model which has 12 layers of transformer blocks with each block having a hidden state size of 768 and 12 multi-head attentions. We trained on four 2080Ti GPUs with a total batch size of 256 for 20 epochs. We used Adam optimizer with an initial learning rate of 0.001. We used a decay learning rate schedule with a warm-up to train the model.

Parameters: We used the following search spaces to adjust the parameters of each modal (bold indicates the choice of the final model): VGG-16 embedding dimensions: {128, 256, **512**, 1024}, GRU hidden states: {50, 100, **128**, 150, 200}, KG image embedding dimensions: {128, 256, **512**, 1024}, $\bar{x}$ dimensions: {50, 100, 150, **200**, 250}, λ_1: {0.1, 0.2, 0.3, **0.4**, 0.5, 0.6, 0.7, 0.8, 0.9} and λ_t: {0.1, 0.2, 0.3, 0.4, 0.5, **0.6**, 0.7, 0.8, 0.9}. We optimize the parameter by Adam with batch size = 10, learning rate = 0.001, β_1 = 0.9, β_2 = 0.999 and epsilon = 10^{-8}.

4.2 Baselines

Because our proposed task is relatively novel, the related models available for our comparison are especially limited. We report the performance of the following state-of-the-art entity linking and visual objects recognition methods as baselines, as well as several configurations of our proposed method to examine contributions of each component (T: textual, V: visual, and KG: knowledge graph).

- Huawei API[⑥] and Tencent API[⑦] (V and KG) use a deep neural network model to recognize people.

⑥ https://www.huaweicloud.com/product/roc.html

⑦ https://cloud.tencent.com/document/api/865/36900

- CoAtt [39] (T and KG) uses a type-aware co-attention model for entity disambiguation.
- Falcon [40] (T and KG) performs joint entity linking of a short text by leveraging several fundamental principles of English morphology.
- CDTE [41] (T and KG) proposes a neural, modular entity linking method using multiple sources of information for entity linking.
- GENRE [42] (T and KG) system realizes entity retrieval by generating entity names. GENRE generates entity names in a token-by-token auto-regressive manner from left to right, and the generated results are affected by context.
- DZMNED [23] (T, V and KG) uses an attention LSTM model for multi-modal NED task in social media posts.
- (Proposed) Our method (T, V and KG) is the proposed method as described in Figure 2.
- (Proposed) Our method without visual features (T and KG) only uses the textual features extracted from the caption and KG.
- (Proposed) Our method without textual features (V and KG) only uses the visual features extracted from the image and KG.
- (Proposed) Our method with a smaller sized KG (T, V and KG)

4.3 Results

Comparison patterns: First of all, because of the novelty of the visual entity linking, we need to rely on existing models to build comparative experimental methods. Next, we will explain the experimental settings and how to achieve relative fairness through settings under some unbalanced condition of experiments, such as V + KG modalities, V + T modalities, and T + KG modalities.

Intuitively, in V + KG modalities, the visual training data of their recognizer network is the work as same as the KG. These massive image data and the image resources in KG modality are equivalent and have the same effect. In the first two experiments of Table 2, we used V modality to replace V + KG modalities to realize the corresponding experiment.

Table 2. Visual entity linking performance on the VELD dataset of Top-1, 3, 5, 10 accuracy on DBpedia 2016.

		Top-1	Top-3	Top-5	Top-10
V+KG	Huawei API	12.53%	18.49%	20.46%	22.94%
V+KG	Tencent API	11.79%	16.42%	21.64%	24.61%
T+KG	Faster-RCNN+CoAtt	55.45%	63.76%	66.05%	67.91%
T+KG	Faster-RCNN+Falcon	56.16%	61.47%	62.17%	63.94%
T+KG	Faster-RCNN+CDTE	58.27%	64.79%	65.09%	66.14%
T+KG	GENRE	73.27%	76.91%	78.84%	81.14%
V+T+KG	DZMNED	66.46%	73.16%	81.06%	83.49%
V+T+KG	**Our method**	**83.16%**	**88.61%**	**92.49%**	**93.81%**

Note: Bounding boxes generation method: N/A or Faster-RCNN.

V + T modalities cannot output the result defined in the task because of lacking the KG entity links. Ignore the KG modality, the target entities, the entity linking task will not continue, so it cannot be used as a comparative experiment. Therefore, in our experiments, due to the lack of target entities, we did not choose the corresponding V + T modalities for comparative analysis.

The short-text entity linking based on the KG (T + KG modalities) cannot link the KG entities to the corresponding image bounding box. For comparison, we first used the scene graph method to generate the corresponding bounding boxes, and then randomly connected the entity in the candidate list to the entity bounding boxes. At the same time, we multiplied each accuracy rate by the number of candidate entities per entity bounding box to ensure the fairness of the accuracy rate. By multiplying the accuracy of the visual entity linking in T + KG modalities by the number of candidate entities, we eliminated the error caused by the random connection of candidate entities in T + KG modalities experiments.

Main results: Table 2 shows the Top-1, 3, 5, and 10 candidate entity list retrieval accuracy results on the VELD dataset. The first two experiments use the information of visual modality and knowledge graph modality. Through the experimental results, we proved that the existing deep neural network based on static off-line training cannot complete the task of visual entity linking well. Because of the limitation of the training dataset, it is difficult to build a dataset which contains image resources of all the entity in the open domain, so the validity of our model is proved from another side.

The third to the fifth experiments are based on the features of textual modality and knowledge graph modality for visual entity linking, and through a series of post-processing, the linking of the target frame is not affected by the visual features. From the experimental results, there is still a large gap between textual modality and our full model.

Compared with the simple visual object recognition methods and textual entity linking method which uses text and KG as the support, we found that our proposed method significantly outperforms these baselines. The reason is that we jointly fused three kinds of features in different modalities, rather than simple modality based linking. Another convincing point is that by applying the similar multi-modal learning model DZMNED on the VELD dataset, the results show that they only achieved 66.46% on the Top-1 accuracy measure. Our model reached 83.16%, which shows that our model has a great advantage in the task of visual entity linking.

Visualization of modality attention: Figure 3 visualizes the modality attention module of our model, where we list each entity (each column) of some samples in the test, in which amplified modality is represented by a darker color, and attenuated modality is shown by a lighter color. We intuitively analyze from the experimental results that more relevant modalities in the visual entity linking will be emphasized through the modality attention module. Specifically, we used the alignments between different modalities from the test set of the VELD dataset.

For our multi-modal visual entity linking model (V+T+KG modalities), we confirm from the experimental results that the modality attention module has successfully enhanced the function of relevant modal

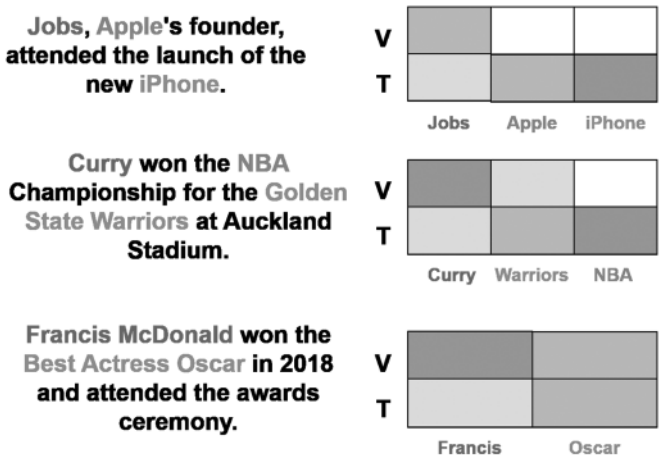

Figure 4. Visualization of modality attention from VELD test data. The model makes final predictions based on the weighted signal from all the modalities. Modalities-V: Visual; T: Textual.

information (e.g., in similar celebrity entity linking), and amplified relevant modality-based contexts in prediction.

In the example of the first row in Figure 4, "Jobs, Apple's founder, attended the launch of the new iPhone", we first generated the candidate entity list of "Jobs", "Apple", "iPhone". For the first example, in the process of entity linking for "Jobs" entity, we learn that the influence of visual modality is higher than that of textual modality according to the color depth of the modalities. For the other two entities "Apple" and "iPhone", the influence of visual modality is much lower than that of textual modality. Because there are few candidate entities of "Apple" and "iPhone", just relying on the textual modality we can easily find the knowledge graph entity corresponding to the contextual semantics, but there are many related entities for "Jobs" entity, so we need to use the feature vector of visual modality for the entity linking task, which is why different entity categories have different modality weights.

In the second example, "Curry won the NBA Championship for the Golden State Warriors at Auckland Stadium.", the candidate entity list is composed of "Curry", "Golden State Warriors" and "NBA". For "Curry" entity, we found that the modality attention module mainly concentrates on visual modality information. For "Golden State Warriors" entity, the proportion of visual modality information and textual modality information is roughly equal, while for "NBA" entity, it mainly depends on textual modality signals.

In the third case, for the person category like "Francis", the modality attention successfully focuses on the visual modality, and attenuates distracting signals, and for "Oscar", visual modality information and textual modality information have the same status.

Ablation study: To evaluate the effectiveness of our different modules, we considered several ablation experiments in Table 3. We validated the effect of the feature in three modalities, visual, textual, and KG.

Because our experimental result is a comprehensive expression of multiple modalities, i.e., the basis of the visual entity linking is an image bounding box, and the caption description generates the candidate entity list, we used the KG entities for links. Therefore, our input data is not changed, and only the parameters of the corresponding part (λ_1 and λ_2) are adjusted to zero in the confidence calculation to achieve the purpose of eliminating a particular modality feature. For the less knowledge graph, we choose the KG embeddings learned from the 1M KG subset. The corresponding experiment results are shown in Table 3.

Table 3. Ablation experiments on our model.

	Accuary	
	Top-1	Top-10
Ours w/o visual features	56.19%	63.42%
Ours w/o textual features	72.55%	82.23%
Ours with a smaller sized KG	60.19%	66.47%
Ours All	**83.16%**	**93.81%**

From the experimental results, it can be found that the features of each modality contribute a certain amount to the performance of visual entity linking. The lack of any modality feature will significantly reduce the performance in terms of accuracy metric. The lack of visual features reduced the top-1 accuracy to 56%. For the absence of text features, the top-1 accuracy decreased to 73%. For reducing the scale of KG, the top-1 accuracy decreased to nearly 60%. These results also suggest that jointly utilizing multi-modal features can obtain the best experimental linking results.

Error analysis: In the example that "Robert Downey Jr. plays Iron Man in the movie", our model links the image bounding boxes to the actor Robert Downey Jr., and ground-truth links it to Iron Man. It means that our model sometimes outputs error results in the situation where one person has multiple roles in KG. Therefore, in some instances, we need to set some rules and constraints to obtain a better result. In other scenarios, when there is occlusion or concealment in the image, there will be deviations. For example, in the second case, we can easily link Suárez in KG to the image bounding box, but it is much more difficult to link Messi because there is a visual occlusion in the image. Such a reason for errors can be attributed to the incompleteness of the image information, and from another aspect, it also indicates the importance of image features for visual entity linking.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a new task called visual entity linking, which links the KG entities to the corresponding image bounding boxes, and we addressed the problem by a novel framework. The proposed framework first extracts the features from three modalities (visual, textual and KG). Then, a deep modal-attention neural network is employed for linking the entities to the corresponding image bounding boxes. We constructed a new dataset VELD for visual entity linking experiments. The experimental results show

that our model achieved state-of-the-art results. Moreover, through extensive ablation experiments, we demonstrated the efficacy of our method.

In the future, a possible improvement direction is to utilize the structural embedding features of the scene graph to improve the performance of visual entity linking. In addition, we hope our model becomes a generic framework for the visual entity linking task, but constructing an ideal complete KG including all the entities in the world is impossible. Therefore, the requirements and effects of visual entity linking need to be determined according to specific applications.

AUTHOR CONTRIBUTIONS

Q.S. Zheng (qiushuo_zheng@seu.edu.cn) led the writing of the paper. He was the leading contributor for task definition, model training, and model tuning. W. Hao (wenhao7841@seu.edu.cn) was the leading contributor for data collection, data propessing and model training. M. Wang (meng.wang@seu.edu.cn) contributed to motivation proposal, task definition and paper revision. G.L. Qi (gqi@seu.edu.cn) proposed the research idea, contributed to motivation proposal and model tuning, and revised the paper.

REFERENCES

[1] Johnson, J., et al.: Image retrieval using scene graphs. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 3668–3678 (2015)
[2] Shen, W., et al.: Linden: Linking named entities with knowledge base via semantic knowledge. In: International World Wide Web Conference, pp. 449–458 (2012)
[3] Lin, T., et al.: Entity linking at Web scale. In: The Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 84–88 (2012)
[4] Nguyen, T.H., et al.: Joint learning of local and global features for entity linking via neural networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2310–2320 (2016)
[5] Chen, Z., Ji, H.: Collaborative ranking: A case study on entity linking. In: Empirical Methods in Natural Language Processing, pp. 771–781 (2011)
[6] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716 (2007)
[7] Pan, X., et al.: Unsupervised entity linking with abstract meaning representation. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1130–1139 (2015)
[8] Chisholm, A., Hachey, B.: Entity disambiguation with Web links. Transactions of the Association for Computational Linguistics 3, 145–156 (2015)
[9] He, Z., et al.: Efficient collective entity linking with stacking. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 426–435 (2013)
[10] Cheng, X., Roth, D.: Relational inference for Wikification. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1787–1796 (2013)
[11] He, K., et al.: Deep residual learning for image recognition. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 770–778 (2016)

[12] Gorniak, P., Roy, D.: Grounded semantic composition for visual scenes. Journal of Artificial Intelligence Research 21, 429–470 (2004)
[13] Lu, C., et al.: Visual relationship detection with language priors. In: European Conference on Computer Vision, pp. 852–869 (2016)
[14] Yao, B., Li, F.-F.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17–24 (2010)
[15] Farhadi, A., et al.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29 (2010)
[16] Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
[17] Xu, D., et al.: Scene graph generation by iterative message passing. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 5410–5419 (2017)
[18] Ramanathan, V., et al.: Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1100–1109 (2015)
[19] Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1745–1752 (2011)
[20] Baltrusaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (2018)
[21] Cao, N.D., et al.: Autoregressive entity retrieval. In: The 9th International Conference on Learning Representations (ICLR 2021). Avaialble at: https://openreview.net/forum?id=5k8F6UU39V/. Accessed 11 December 2021
[22] Zhu, Y., et al.: Knowledge perceived multi-modal pretraining in e-commerce. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2744–2752 (2021)
[23] Moon, S., Neves, L., Carvalho, V.: Multimodal named entity disambiguation for noisy social media posts. In: Annual Meeting of the Association for Computational Linguistics, pp. 2000–2008 (2018)
[24] Biten, A.F., et al.: Good news, everyone! context driven entity-aware captioning for news images. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 12466–12475 (2019)
[25] Lu, D., et al.: Entity-aware image caption generation. arXiv preprint arXiv:1804.07889 (2018)
[26] Antol, S., et al.: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
[27] Tsai, Y.-H.H., et al.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference Association for Computational Linguistics, pp. 6558–6569 (2019)
[28] Sil, A., et al.: Multi-lingual entity discovery and linking. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 22–29 (2018)
[29] Li, M., et al.: Gaia: A finegrained multimedia knowledge extraction system. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 77–86 (2020)
[30] Zhang, J., et al.: Graphical contrastive losses for scene graph parsing. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 11535–11543 (2019)
[31] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
[32] Chung, J., et al.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
[33] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[34] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
[35] Auer, S., et al.: DBpedia: A nucleus for a Web of open data. In: International Semantic Web Conference & Asian Semantic Web Conference, pp. 722–735 (2007)
[36] Trouillon, T., et al.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning, pp. 2071–2080 (2016)
[37] Ramisa, A., et al.: Article annotation by image and text processing. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(5), 1072–1085(2017)
[38] Lin, T.-Y., et al.: Microsoft coco: Common objects in context. In: Europeon Conference on Computer Vision, pp. 740–755 (2014)
[39] Nie, F., et al.: Mention and entity description co-attention for entity disambiguation. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018), pp. 5908–5915 (2018)
[40] Sakor, A., et al.: Old is gold: Linguistic driven approach for entity and relation linking of short text. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2336–2346 (2019)
[41] Gupta, N., Singh, S., Roth, D.: Entity linking via joint encoding of types, descriptions, and context. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2681–2690 (2017)
[42] De Cao, N., et al.: Autoregressive entity retrieval. arXiv preprint arXiv:2010.00904 (2021)

AUTHOR BIOGRAPHY

Qiushuo Zheng is a graduate student at School of Cyber Science and Engineering, Southeast University. He received a bachelor's degree from Southeast University. His main research interests are multi-modal learning and downstream applications of knowledge graph.
ORCID: 0000-0003-4921-2218

Hao Wen is an undergraduate student at the School of Computer Science and Engineering, Southeast University. Currently, his research interests mainly include information retrieval, entity linking and multi-media research.
ORCID: 0000-0003-3165-3859

Meng Wang is an Assistant Professor in the Knowledge Graph & AI Research Group, School of Computer Science and Engineering, Southeast University (SEU), China. He is also a SEU Zhishan Young Scholar. He obtained his doctoral degree from the Department of Computer Science and Technology, Xi'an Jiaotong University in 2018, under the supervision of Prof. Jun Liu. He was a visiting scholar, working with Prof. Xue Li and Prof. Xiaofang Zhou, in the DKE Lab at University of Queensland, Australia in 2016. His research area is in the knowledge graph, semantic search, natural luangauge processing (NLP), and cross-modal data.
ORCID: 0000-0002-2293-1709

Guilin Qi is a Professor at Southeast University, China, where he also serves as Director of the Institute of Cognitive Intelligence of Southeast University. He was supported by the Six Talent Peak Programs of Jiangsu Province. At present, he is the Deputy Director of the Language and Knowledge Computing Professional Committee of Chinese Information Society and the Deputy Director of the Knowledge Organization Professional Committee of China Science and Technology Information Society. He was a visiting professor at Griffith University in Australia and a visiting professor at the First University of Toulouse in France. He graduated from Yichun University, majoring in mathematics, in 1998, obtained a Master's degree from the Mathematics and Information Department of Jiangxi Normal University in 2002 and a Doctor's degree in Computer Science from Queen's University of Belfast in 2006.
ORCID: 0000-0002-1957-6961